The construction of Twitter databases.
Empirical case studies on the socio-technical meaning of Twitter data as a research tool

Evelien D'heer, iMinds – MICT – Ghent University

Pieter Verdegem, iMinds – MICT – Ghent University

This paper deals with methodological challenges related to Twitter research. In particular we focus on (1) unfound users and deleted tweets (that resurrect), (2) URLs that do not link (correctly) and (3) the limits of hashtag samples to study conversations. The empirical case studies we present are part of a larger research project on social media, elections and public debate. These issues are not unique for our data, but are of general relevance for anyone working with Twitter data.

Departing from the idea that a database is "anything but a simple collection of items" (Manovich, 2001, p. 194), we scrutinize the way APIs deliver and *structure* data. Based on our case studies, we understand datasets as *textual* representations of user activity (e.g. images are stored as URLs), presented in *chronological* rather than "conversational" order. In addition, whereas data collection is *real-time*, the manual analysis of the data often is not, resulting in unidentifiable users and tweets. Last, APIs provide "exact matches" for our hashtag-based data requests. However, when we include non-hashtagged responses, we notice the hashtag approach systematically underestimates reciprocity between users.

We departed from a selection of empirical cases to understand Twitter data(bases) as constructions. In general, awareness on the construction of Twitter data is crucial, as we build upon this data to explain socio-cultural phenomena.

Reference:

Manovich, L. (2001) The language of new media. MIT Press: Cambridge.